**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KỲ**

**MÔN HỌC: NHẬP MÔN LẬP TRÌNH PYTHON**

**QUẢN LÝ DANH SÁCH KHÁCH HÀNG**

**CỦA MỘT CỬA HÀNG Ô TÔ**

**VÀO THÁNG 12 NĂM 2023**

**Mã lớp học phần: IPPA233277**

**Học kỳ 1 – Năm học 2024-2025**

**Giảng viên hướng dẫn: ThS. Trần Quang Khải**

**Danh sách sinh viên thực hiện:**

|  |  |
| --- | --- |
| **MSSV** | **Họ tên** |
| 23110202 | Nguyễn Quốc Đạt |
| 23110361 | Nguyễn Huỳnh Tự |
| 23110294 | Nguyễn Vũ Quân |

***Thành phố Hồ Chí Minh, tháng 11 năm 2024***

*Nhận xét của giảng viên*

*TP. Hồ Chí Minh, ngày … tháng… năm 2024*

*Giảng viên ký tên*

**MỤC LỤC**

[**PHẦN 1: MỞ ĐẦU 1**](#_Toc14341)

[**1. Lý do chọn đề tài 1**](#_Toc675)

[**2. Mục tiêu đề tài 1**](#_Toc12201)

[**3. Đối tượng và phạm vi nghiên cứu 1**](#_Toc2476)

[**PHẦN 2. NỘI DUNG 3**](#_Toc337)

[**CHƯƠNG 1: TÓM TẮT CƠ SỞ LÝ THUYẾT 3**](#_Toc20101)

[**1.1. Giới thiệu về ngôn ngữ Python 3**](#_Toc15093)

[**1.2. Giới thiệu về các thư viện thường được sử dụng trong Python và môi trường lập trình 3**](#_Toc26672)

[**CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU VỀ KHÁCH HÀNG CỦA CỬA HÀNG XE Ô TÔ VÀO THÁNG 12 NĂM 2023 5**](#_Toc31946)

[**2.1. Các tập dữ liệu cần thu thập 5**](#_Toc20042)

[**CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU KHÁCH HÀNG CỦA CỬA HÀNG Ô TÔ VÀO THÁNG 12 NĂM 2023 …….7**](#_Toc16207)

[**3.1. Khởi tạo dữ liệu 7**](#_Toc16737)

[**3.1.1. Import các thư viện cần thiết 7**](#_Toc27326)

[**3.1.2. Xây dựng đường dẫn đến tập tin data.csv 7**](#_Toc4892)

[**3.2. Làm sạch dữ liệu (Data validation and cleaning) 8**](#_Toc21261)

[**3.3. Mô tả thống kê tập dữ liệu data.csv 13**](#_Toc13512)

[**3.4. Phân tích dữ liệu (Data analysis) 13**](#_Toc1248)

[**3.4.1. Phân tích các thương hiệu xe bán chạy nhất của cửa hàng ô tô 13**](#_Toc26355)

[**3.4.2. Phân tích doanh thu của các hãng xe bán chạy 15**](#_Toc29464)

[**3.4.3. Phân tích các nhân viên có thành tích nổi bật của cửa hàng ô tô 18**](#_Toc31553)

[**3.4.4. Phân tích xu hướng màu thích yêu thích của khách hàng lựa chọn 19**](#_Toc20971)

[**CHƯƠNG 4: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN, KIỂM CHỨNG VÀ ĐÁNH GIÁ MÔ HÌNH CHO TẬP DỮ LIỆU CỬA HÀNG XE Ô TÔ VÀO THÁNG 12 21**](#_Toc25002)

[**4.1. Tổng quan dữ liệu (Data Overview) 21**](#_Toc5031)

[**4.2. Trực quan hóa vị trí taxi đón khách sử dụng plotly express 23**](#_Toc17584)

[**4.3. Trực quan hóa các tuyến đường taxi sử dụng plotly graph objects 24**](#_Toc23688)

[**4.4. Tiền xử lý dữ liệu (Preprocessing Data) 25**](#_Toc21592)

[**4.4.1. Tính khoảng cách Haversine (Compute Haversine distance) 25**](#_Toc11853)

[**4.4.2. Tính khoảng cách radius (Compute Radius Distance) 26**](#_Toc7664)

[**4.4.3. Tính khoảng cách đến các địa điểm nổi bật ở New York (Add Airport Locate) 26**](#_Toc20850)

[**4.4.4. Tiền xử lý (Preprocess) 27**](#_Toc32477)

[**4.4. Phân tích đặc trưng (Feature Analytic) 28**](#_Toc28982)

[**4.5. Xây dựng Model (Build Model) 29**](#_Toc29452)

[**4.5.1. Train-Val split 29**](#_Toc18648)

[**4.5.2. Train Model 31**](#_Toc17339)

[**4.6. Mô hình Eval (Model Eval) 31**](#_Toc29614)

[**PHẦN KẾT LUẬN 33**](#_Toc434)

[**PHẦN TÀI LIỆU THAM KHẢO 34**](#_Toc16131)

**DANH MỤC HÌNH ẢNH**

[Hình 1 . Hình 1. Tập dữ liệu data.csv lưu trữ thông tin các khách hàng của một cửa hàng ô tô vào tháng 12 năm 2023 5](#_Toc2196)

.[Hình 2 . Code import các thư viện cần thiết cho việc phân tích và xây dự mô hình quản lý khách hàng của cửa hàng ô tô vào tháng 12 năm 2023. 7](#_Toc24905)

[Hình 4 . Code xây dựng đường dẫn đến tập tin train.csv và test.csv](#_Toc11487)

[trên máy tính 7](#_Toc11487)

[Hình 5 . Kết quả khi chạy code xây dựng đường dẫn 8](#_Toc9857)

[Hình 6 . Code xây dựng đường dẫn đến tập tin train.csv và test.csv trên Kaggle 8](#_Toc19902)

[Hình 7 . Code kiểm tra tập dữ liêu train.csv 9](#_Toc15446)

[Hình 8 . Một số kết quả khi kiểm tra tập dữ liệu train.csv 9](#_Toc1386)

[Hình 9 . Một số kết quả khi kiểm tra tập dữ liệu train.csv 10](#_Toc31490)

[Hình 10 . Kiểm nghiệm những dòng có giá trị pickup\_longtitude >0 11](#_Toc29961)

[Hình 11 . Code tạo các hàm để clean dữ liệu 12](#_Toc19697)

[Hình 12 . Code của việc mô tả thống kê tập dữ liệu train.csv 13](#_Toc7062)

[Hình 13 . Kết quả của việc mô tả thống kê tập dữ liệu train.csv 13](#_Toc29467)

[Hình 14 . Code tạo biểu đồ phân phối giá cước taxi 13](#_Toc23767)

[Hình 15 . Biểu đồ phân phối giá cước taxi 14](#_Toc25488)

[Hình 16 . Code tạo biểu đồ phân tích trung bình giá cước taxi](#_Toc29839)

[theo ngày trong tuần 15](#_Toc29839)

[Hình 17 . Biểu đồ trung bình giá cước taxi theo ngày trong tuần 15](#_Toc30292)

[Hình 18 . Code tạo biểu đồ phân tích trung bình giá cước taxi](#_Toc10742)

[theo giờ trong ngày 16](#_Toc10742)

[Hình 19 . Biểu đồ trung bình giá cước taxi theo giờ trong ngày 17](#_Toc21614)

[Hình 20 . Code vẽ biểu đồ phân tán phân tích vị trí đón khách 18](#_Toc22374)

[Hình 21 . Biểu đồ phân tích vị trí taxi đón khách và giá cước 18](#_Toc23868)

[Hình 22 . Code tạo biểu đồ đường phân tích xu hướng giá cước taxi](#_Toc6743)

[qua từng năm 19](#_Toc6743)

[Hình 23 . Biểu đồ trung bình giá cước taxi qua từng năm 19](#_Toc30217)

[Hình 24 . Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra](#_Toc16269)

[và clean data 21](#_Toc16269)

[Hình 25 . Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra](#_Toc30689)

[và clean data 22](#_Toc30689)

[Hình 26 . Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra](#_Toc16272)

[và clean data 23](#_Toc16272)

[Hình 27 . Code tạo bản đồ trực quan hóa các vị trí taxi đón khách](#_Toc12196)

[ở New York City 23](#_Toc12196)

[Hình 28 . Bản đồ các vị trí mà taxi đón khách ở New York City 24](#_Toc24413)

[Hình 29 . Code kết hợp các cột liên quan từ hai DataFrame train và test 24](#_Toc9306)

[Hình 30 . Code tạo bản đồ trực quan hóa các tuyến đường taxi](#_Toc18779)

[ở New York City 24](#_Toc18779)

[Hình 31 . Bản đồ các tuyến đường taxi ở New York City 25](#_Toc6086)

[Hình 32 . Code tạo hàm tính khoảng cách Haversine 25](#_Toc12944)

[Hình 33 . Code tạo hàm để tính toán hướng (bearing) 26](#_Toc30014)

[Hình 34 . Code tạo hàm tính khoảng cách đến các địa điểm nổi bật](#_Toc24597)

[ở New York City 27](#_Toc24597)

[Hình 35 . Code tạo các hàm tiền xử lý dữ liệu 28](#_Toc8890)

[Hình 36 . Code thực hiện phân tích và tiền xử lý dữ liệu 28](#_Toc8391)

[Hình 37 . Code trực quan hóa phân phối đặc trưng trong dữ liệu huấn luyện 29](#_Toc26666)

[Hình 38 . Kết quả trực quan hóa phân phối đặc trưng trong dữ liệu huấn luyện 30](#_Toc3904)

[Hình 39 . Code chuẩn bị dữ liệu huấn luyện và validation](#_Toc24845)

[cho mô hình LightGBM 30](#_Toc24845)

[Hình 40 . Code huấn luyện mô hình 31](#_Toc12294)

[Hình 41 .Đánh giá hiệu suất mô hình 32](#_Toc11327)

**DANH MỤC BẢNG**

[Bảng 1 . Bảng mô tả tập dữ liệu data.csv 5](#_Toc150)

[Bảng 2 . Bảng mô tả tập dữ liệu train.csv 6](#_Toc27695)

**BẢNG DANH MỤC CÁC TỪ VIẾT TẮT**

|  |  |
| --- | --- |
| **Ký hiệu chữ viết tắt** | **Cách viết đầy đủ** |
| NYC | New York City |
| JFK | John F. Kennedy International Airport |
| EWR | Newark Liberty International Airport |
| LGA | LaGuardia Airport |

**KẾ HOẠCH PHÂN CÔNG NHIỆM VỤ THỰC HIỆN ĐỀ TÀI**

**CUỐI KỲ MÔN LẬP TRÌNH PYTHON**

**HỌC KỲ I NĂM HỌC 2023-2024**

**1. Mã lớp môn học:** **IPPA233277\_06**

**2. Giảng viên hướng dẫn: ThS. Trần Quang Khải**

**3. Tên đề tài: QUẢN LÝ DANH SÁCH XE Ô TÔ CỦA MỘT CỬA HÀNG VÀO THÁNG 12 NĂM 2023**

**4. Bảng phân công nhiệm vụ:**

|  |  |
| --- | --- |
| **Sinh viên thực hiện** | **Nội dung thực hiện** |
| Nguyễn Quốc Đạt | * Code xử lý dữ liệu và phân tích dữ liệu. * Xây dựng báo cáo Phần Mở đầu và Phần Nội Dung (Chương 1, 2, 3). * Tổng hợp và hoàn thiện báo cáo. |
| Nguyễn Vũ Quân | * Code xử lý dữ liệu và xây dựng, huấn luyện mô hình. * Xây dựng báo cáo Phần Kết Luận và Phần Nội Dung (Chương 4). |
| Nguyễn Huỳnh Tự |  |

**LỜI CẢM ƠN**

Để thực hiện được đề tài này, em xin gửi lời cảm ơn chân thành đến thầy Trần Quang Khải - giảng viên dạy học và hướng dẫn bộ môn Học Máy của lớp IPPA233277 vì đã chỉ dẫn và hỗ trợ chúng em hết mình trong việc học tập và trong quá trình thực hiện tiểu luận cuối kỳ.

Trong quá trình học tập, nghiên cứu và thực hiện đề tài chúng em đã nổ lực và cố gắng rất nhiều, nhưng do kinh nghiệm và hiểu biết còn hạn hẹp nên có thể phần báo cáo có thể còn nhiều thiếu sót. Chúng em rất mong nhận được sự động viên, hỗ trợ và góp ý từ thầy để nhóm em có thể hoàn thiện hơn không chỉ trong môn học này mà còn làm tốt hơn nữa trong những chặng đường sắp tới.

Một lần nữa, chúng em xin trân trọng cảm ơn sự quan tâm giúp đỡ của thầy trong suốt hành trình vừa qua.

# PHẦN 1: MỞ ĐẦU

## Lý do chọn đề tài

Quản lý khách hàng là một trong những yếu tố quan trọng quyết định sự thành công của một doanh nghiệp, đặc biệt là trong ngành công nghiệp ô tô – nơi mà sự cạnh tranh ngày càng gay gắt. Đối với một cửa hàng ô tô, việc hiểu rõ nhu cầu, thói quen mua sắm, và hành vi của khách hàng không chỉ giúp cải thiện trải nghiệm dịch vụ mà còn tăng cường khả năng giữ chân khách hàng và thu hút thêm nhiều khách hàng tiềm năng.

Cửa hàng ô tô thường phải đối mặt với những thách thức như quản lý lượng lớn thông tin khách hàng, theo dõi lịch sử mua bán, và tối ưu hóa các chiến dịch tiếp thị. Một hệ thống quản lý khách hàng hiệu quả có thể giúp cửa hàng giải quyết những vấn đề này bằng cách tổ chức thông tin một cách khoa học, hỗ trợ ra quyết định nhanh chóng và chính xác.

Nhóm thực hiện báo cáo lựa chọn đề tài này vì nhận thấy tầm quan trọng của việc xây dựng một hệ thống quản lý khách hàng hiện đại, không chỉ giúp tăng cường hiệu quả kinh doanh mà còn nâng cao chất lượng dịch vụ. Dựa trên những kiến thức đã học từ bộ môn Lập trình cơ bản Python, chúng em mong muốn ứng dụng các công nghệ hiện đại để xây dựng một hệ thống quản lý khách hàng phù hợp, đáp ứng tốt các yêu cầu thực tế và đóng góp vào sự phát triển bền vững của cửa hàng ô tô.

## Mục tiêu đề tài

Mục tiêu của đề tài là xây dựng một hệ thống quản lý khách hàng hiệu quả, giúp cửa hàng ô tô theo dõi, lưu trữ và phân tích thông tin khách hàng một cách khoa học. Hệ thống này sẽ hỗ trợ trong việc quản lý lịch sử mua hàng, lịch bảo trì xe, và nhu cầu cá nhân của từng khách hàng. Nhóm em sẽ ứng dụng các kỹ thuật phân tích dữ liệu để phát triển công cụ hỗ trợ tối ưu hóa quy trình chăm sóc khách hàng, cải thiện chất lượng dịch vụ và tăng cường khả năng giữ chân khách hàng. Hệ thống không chỉ giúp cửa hàng nâng cao hiệu quả kinh doanh mà còn góp phần xây dựng mối quan hệ lâu dài, bền vững với khách hàng.

## Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài này là:

* Tập dữ liệu khách hàng của cửa hàng ô tô vào tháng 12 năm 2023: Bao gồm thông tin về mã số của xe,ngày giao dịch,giới tính khách hàng,thu nhập hằng năm,đại lý bán ra,mẫu xe, công ty sản xuất xe, tên khách hàng và giá bán ra cũng như màu sắc của xe.

Phạm vi nghiên cứu, chúng em sẽ tập trung vào việc:

* Phân tích dữ liệu: Tiến hành khai thác, xử lý dữ liệu lớn từ tệp dữ liệu khách hàng của cửa hàng ô tô.
* Xây dựng mô hình quản lý: Áp dụng các kiến thức về ngôn ngữ python để xây dựng một mô hình, hệ thống quản lý khách hàng một cách chính xác và hiệu quả nhất.

# PHẦN 2. NỘI DUNG

## CHƯƠNG 1: TÓM TẮT CƠ SỞ LÝ THUYẾT

### 1.1. Giới thiệu về ngôn ngữ Python

Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (Machine Learning). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau.

Dưới đây là những định nghĩa khái quát về Python:

Python is a high-level, object-oriented programming language with simple and readable syntax, enabling quick and efficient coding Python là một ngôn ngữ lập trình bậc cao, hướng đối tượng, với cú pháp đơn giản và dễ đọc, cho phép viết mã nguồn nhanh chóng và hiệu quả.)

– Guido van Rossum, 1991

### 1.2. Giới thiệu về các thư viện thường được sử dụng trong Học Máy và môi trường lập trình

Các thư viện phổ biến được sử dụng trong Machine Learning bao gồm:

1. NumPy (Numerical Python)

NumPy cung cấp một đối tượng đa chiều (ndarray) hiệu quả, giúp thực hiện các phép toán số học nhanh chóng. Dùng trong các phép toán ma trận và các thao tác tính toán khoa học

2. Pandas

Pandas cung cấp các cấu trúc dữ liệu linh hoạt DataFarme giúp làm việc với dữ liệu dạng bảng dễ dàng. Đọc được nhiều định dạng file như CSV, SQL, …, làm sạch dữ liệu, thực hiện các thao tác cơ bản trên dữ liệu.

3. Matplotlib

Matplotlib tạo các biểu đồ và đồ thị để hiển thị dữ liệu. Hỗ trợ vẽ nhiều loại biểu đồ như đường, cột, hình trò và nhiều loại biểu đồ khác.

5. Image,ImageTk

Image: Đây là lớp chính cho phép thực hiện các thao tác như mở, chỉnh sửa, và lưu trữ hình ảnh. Với Image, người dùng có thể đọc các tệp hình ảnh từ nhiều định dạng khác nhau như JPEG, PNG, BMP, hoặc GIF, cũng như thực hiện các thao tác xử lý hình ảnh như thay đổi kích thước, cắt, xoay, áp dụng bộ lọc, và chuyển đổi định dạng. Image là công cụ cốt lõi để xử lý dữ liệu hình ảnh trong các ứng dụng Python

ImageTk: Đây là lớp hỗ trợ tích hợp hình ảnh với Tkinter, thư viện giao diện đồ họa phổ biến của Python. ImageTk cho phép chuyển đổi các đối tượng Image thành định dạng mà các widget Tkinter có thể hiển thị, chẳng hạn như thông qua ImageTk.PhotoImage. Điều này rất hữu ích khi xây dựng giao diện đồ họa hiển thị hình ảnh trực tiếp trong các ứng dụng Python.

6. Os

Thư viện cung cấp các chức năng để quản lý và thao tác với các tệp và thư mục, chẳng hạn như tạo, đổi tên, xóa tệp hoặc thư mục, và thay đổi thư mục làm việc hiện tại.

7. Csv

Thư viện csv trong Python là một công cụ chuyên dụng để làm việc với các tệp CSV (Comma-Separated Values), cho phép đọc, ghi, và xử lý dữ liệu dạng bảng một cách dễ dàng và hiệu quả.

## CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY

### 2.1. Các tập dữ liệu cần thu thập

Ở đề tài này, nhóm em đã thu thập những tập dữ liệu có chứa những thông tin chi tiết về các khách hàng của cửa hàng ô tô vào tháng 12 năm 2023:

* File data.csv:
  + Được sử dụng để huấn luyện mô hình.
  + Danh sách các đặc trưng đầu vào bao gồm: **Car\_id** (mã số của khách hàng khi mua xe), **Date** (ngày giao dịch, định dạng mm/dd/yyyy), **Customer Name** (tên khách hàng), **Gender** (giới tính của khách hàng, ví dụ: Male/Female), **Annual Income** (thu nhập hằng năm của khách hàng), **Dealer\_Name** (tên đại lý bán xe), **Company** (tên công ty sản xuất xe), **Model** (mẫu xe), **Color** (màu sắc của xe), **Price** (giá bán của xe) và **Phone** (số điện thoại của khách hàng).
  + Thường có số lượng dòng (samples) lớn hơn để mô hình có đủ dữ liệu để học.

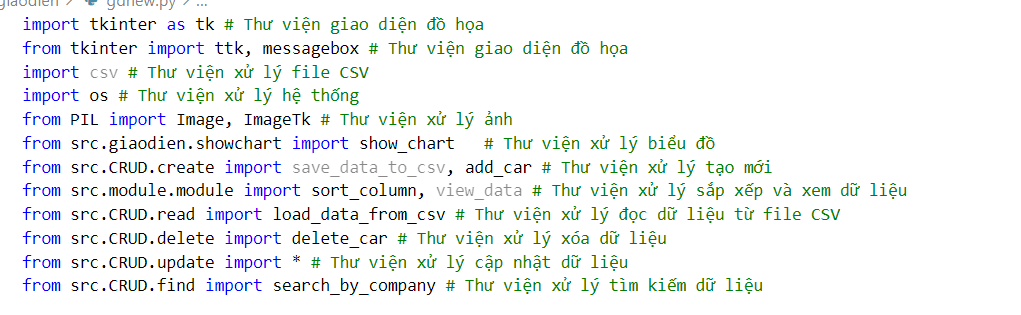
|  |  |
| --- | --- |
| **Thông tin** | **Mô tả** |
| Tên tập dữ liệu | data.csv |
| Kích thước | 11 cột và khoảng 2000 dòng dữ liệu |
| Loại dữ liệu | Số nguyên, số thực, thời gian, chuỗi ký tự |
| *Hình 1**. Tập dữ liệu data.csv lưu trữ thông tin các khách hàng của một cửa hàng ô tô vào tháng 12 năm 2023* | |

*Bảng 1**. Bảng mô tả tập dữ liệu data.csv*

## CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY

### 3.1. Khởi tạo dữ liệu

#### 3.1.1. Import các thư viện cần thiết

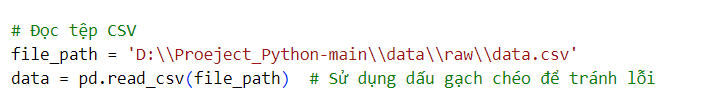


*Hình* *2.* *Code import các thư viện cần thiết cho việc phân tích và xây dự mô hình quản lý khách hàng của cửa hàng ô tô vào tháng 12 năm 2023.*

#### 3.1.2. Xây dựng đường dẫn đến tập tin train.csv và test.csv

Tùy vào thư mục chứa các tập tin này ở đường dẫn nào thì điều hướng sang đường dẫn đó.

* Lưu tập tin trên Kaggle

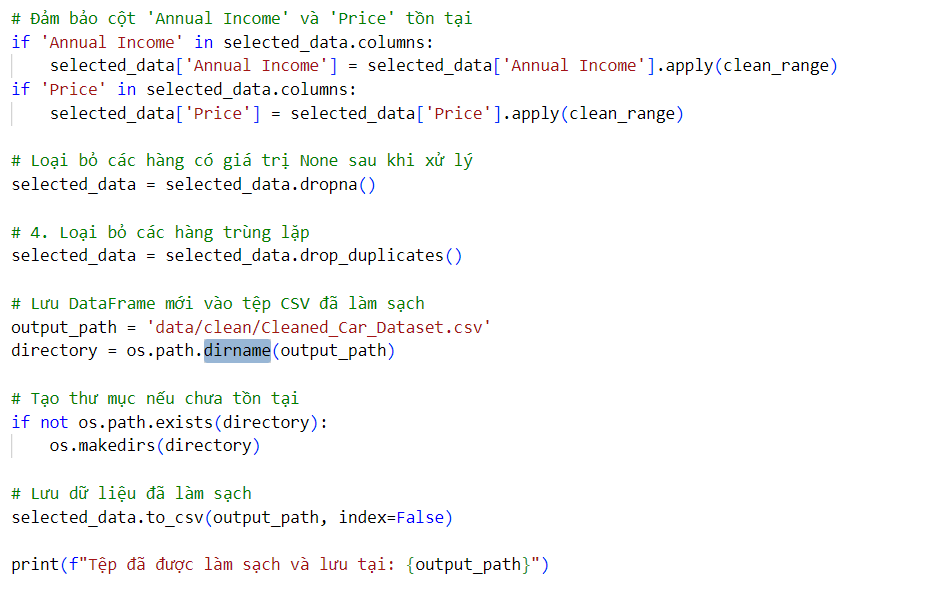


*Hình* *3. Code xây dựng đường dẫn đến tập tin data.csv trên Kaggle*

### 3.2. Làm sạch dữ liệu (Data validation and cleaning)

Ở bước này, chúng ta sẽ kiểm tra xem tập dữ liệu có giá trị null hay có những giá trị nào bị lỗi không và xử lý chúng. Và chúng ta sẽ tiến hành “làm sạch dữ liệu” để loại bỏ các dữ liệu bị lỗi (nếu có).





*Hình* *4. Code tạo các hàm để clean dữ liệu*

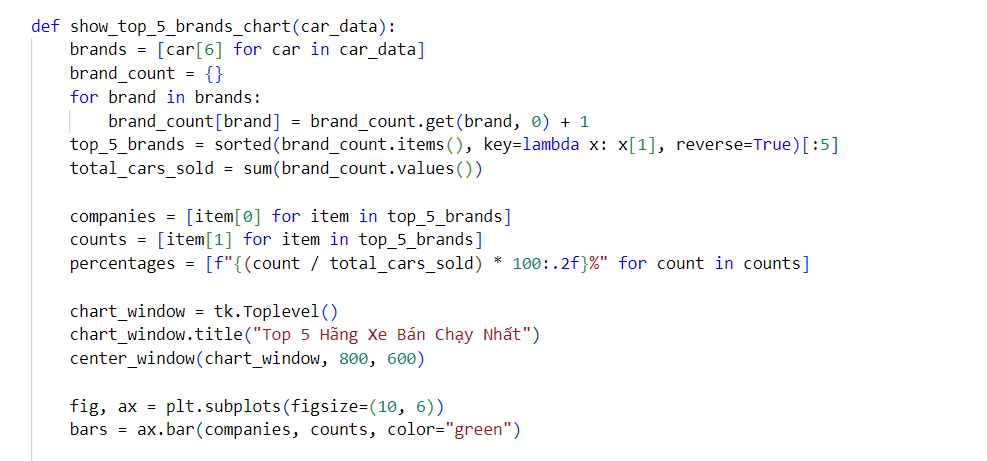
Tóm lại:

* clean\_range: Hàm này làm sạch dữ liệu bằng chuẩn hóa định dạng cột Annual Income và Price, xử lý nếu người dùng nhập khoảng giá trị,giá trị không đúng hoặc bỏ trống giá trị, và loại bỏ các giá trị không hợp lệ.
* dropna: Hàm này loại bỏ các hàng mang giá trị None sau khi chạy hàm clean\_range.
* Drop\_duplicates: Hàm này xóa các giá trị trùng lặp
* Dirname: Hàm này thiết lập nên thư mục theo đường dẫn đã tạo ở trên

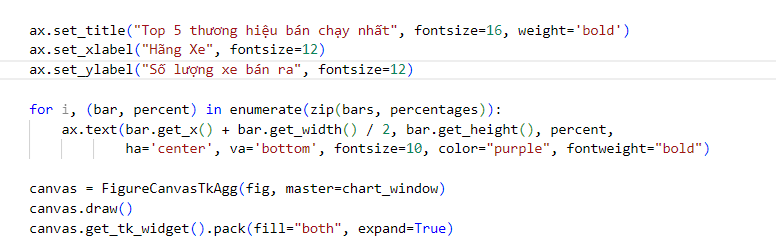
### 3.3. Phân tích dữ liệu (Data analysis)

#### 3.3.1. Phân tích các thương hiệu xe bán chạy nhất của cửa hàng ô tô

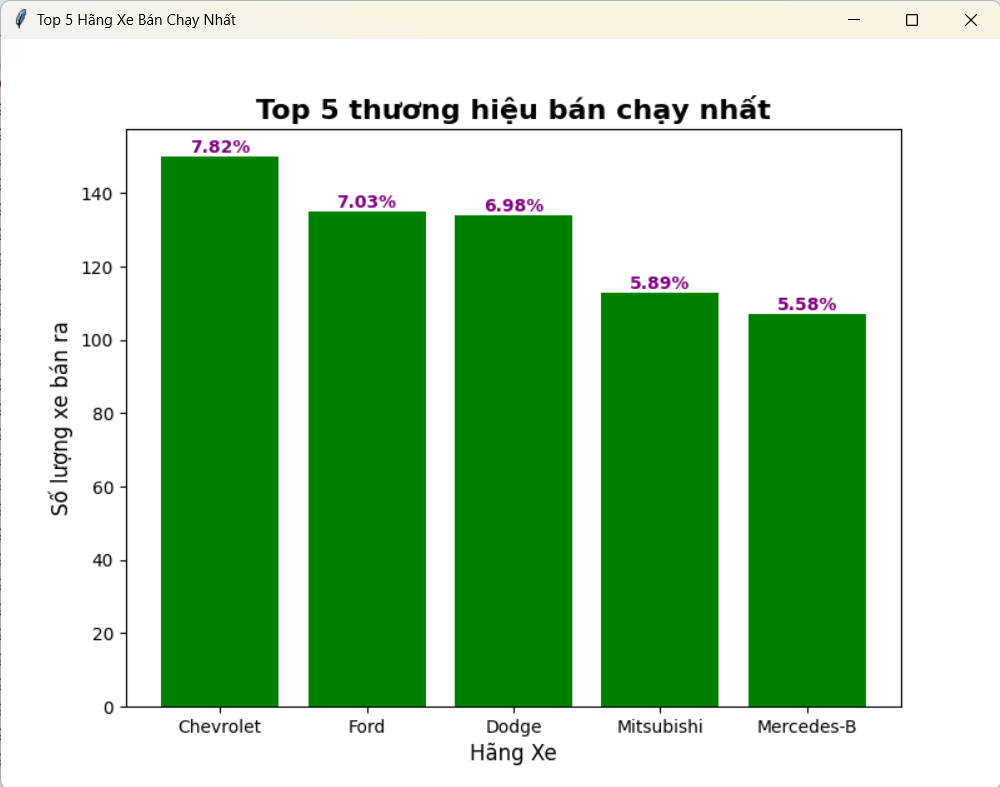
Tại phần này, nhóm báo cáo đồ án sẽ phân tích các thương hiệu xe bán chạy nhất của cửa hàng ô tô để kiểm tra xem đâu là những lựa chọn hàng đầu của các khách hàng khi đến mua xe. Từ đó cửa hàng ưu tiên nhập các thương hiệu xe đó và đảm bảo doanh thu cho cửa hàng luôn gia tăng và phát triển.



*Hình 14**. Code tạo biểu đồ top 5 thương hiệu xe bán chạy nhất của cửa hàng ô tô*



*Hình 14. Code tạo biểu đồ top 5 thương hiệu xe bán chạy nhất của cửa hàng ô tô*



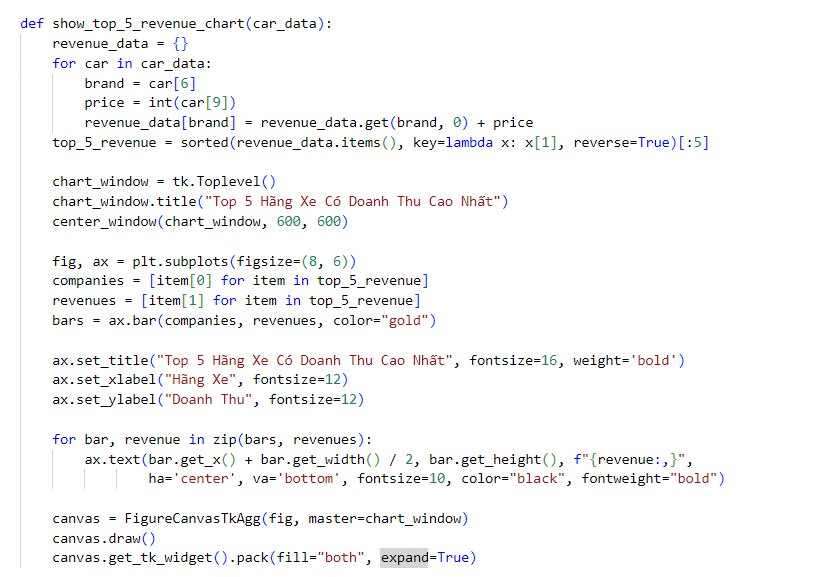
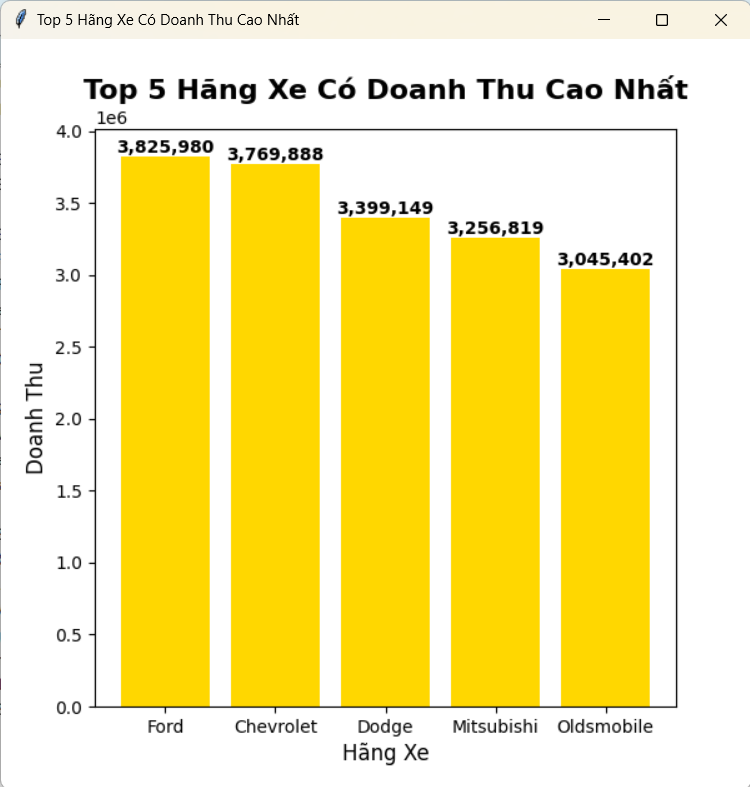
*Hình 15**. Biểu đồ top 5 thương hiệu xe bán chạy nhất của cửa hàng ô tô*

Dựa vào biểu đồ "Top 5 Best-selling Companies – Third Quarter / 2024" ta có thể thấy:

* Phần lớn các hãng xe trên đều bán được hơn 100 chiếc .
* Hãng xe phổ biến nhất chính là Chevrolet chiếm tỉ lệ 7.82%, tiếp đó là Ford với 7.03%, Dodge( 6.98%), Mitsubishi(5.89%), và cuối cùng chính là Mercedes-B với con số 5.58% .

#### 3.3.2. Phân tích doanh thu của các hãng xe bán chạy

Tại đây, chúng ta sẽ phân tích doanh thu của các hãng xe bán chạy.

*Hình 16**.* *Code tạo biểu đồ phân tích doanh thu của các hãng xe bán chạy* 

*Hình 17**. Biểu đồ doanh thu của các hãng xe bán chạy*

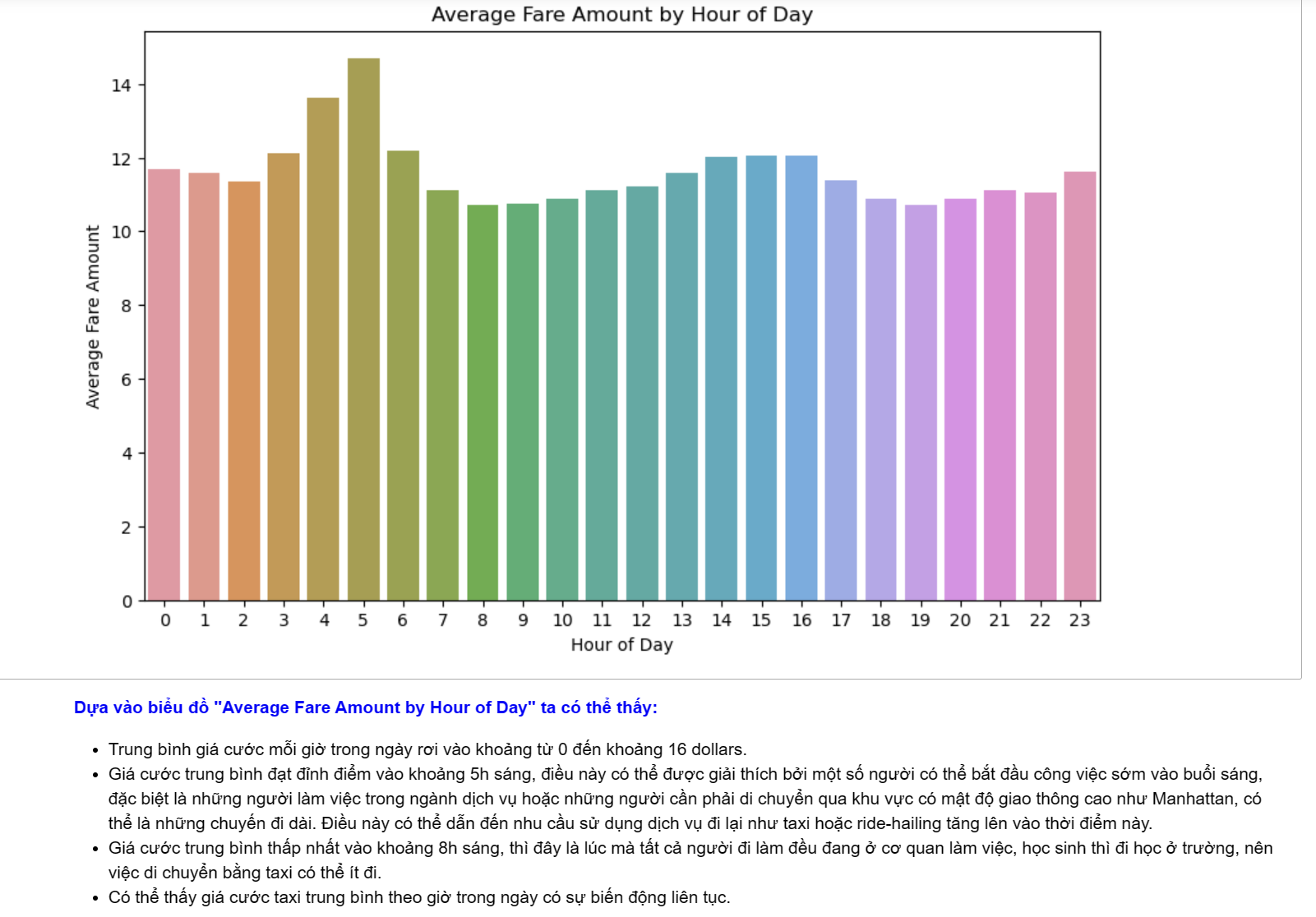
Dựa vào biểu đồ "Top 5 hãng xe có doanh thu cao nhất" ta có thể thấy:

* Ta có thể thấy Ford dẫn đầu với 3.825.980$,Chevrolet (3.769.888$),Dodge (3.399.149), Mitsubishi (3.256.819) và cuối cùng là Oldsmobile (3.045.402).
* Đặc biệt nếu ta liên kết với biểu đồ “top 5 thương hiệu xe bán chạy nhất của cửa hàng ô tô” mặc dù Oldsmobile không nằm trong top những thương hiệu bán chạy nhưng vẫn đạt được một doanh số khủng .
* Ngày thứ 7 thì có trung bình giá cước thấp nhất.
* Các ngày còn lại trong tuần có sự phân bố giá cước trung bình khá đều nhau, có thể đây là do nhu cầu di chuyển bằng taxi khi đi làm của người dân ở New York City.

##### 3.4.2.2. Phân tích trung bình giá cước taxi theo giờ trong ngày



*Hình 18**. Code tạo biểu đồ phân tích trung bình giá cước taxi theo giờ trong ngày*



*Hình 19**. Biểu đồ trung bình giá cước taxi theo giờ trong ngày*

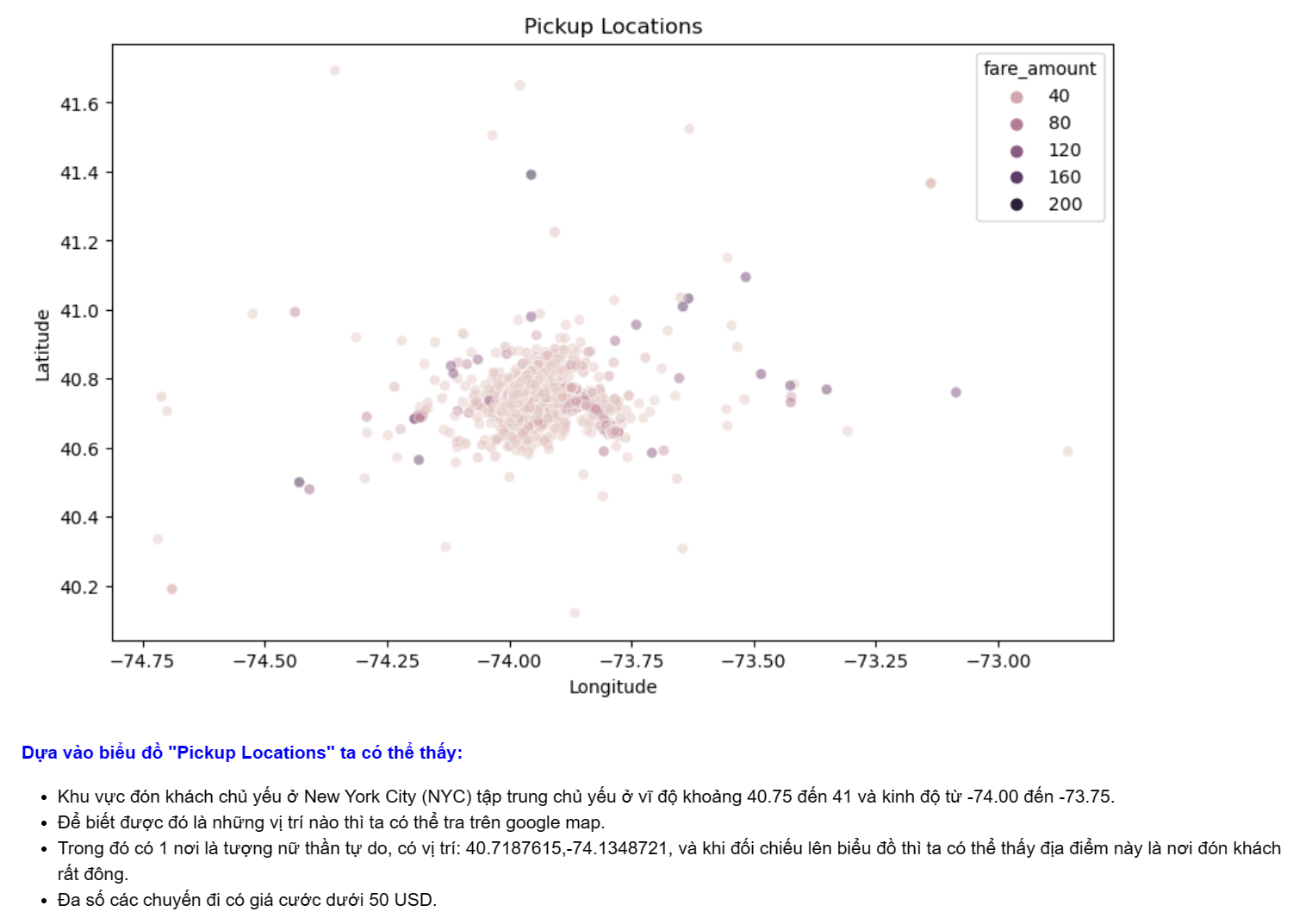
Dựa vào biểu đồ "Average Fare Amount by Hour of Day" ta có thể thấy:

* Trung bình giá cước mỗi giờ trong ngày rơi vào khoảng từ 0 đến khoảng 16 dollars.
* Giá cước trung bình đạt đỉnh điểm vào khoảng 5h sáng, điều này có thể được giải thích bởi một số người có thể bắt đầu công việc sớm vào buổi sáng, đặc biệt là những người làm việc trong ngành dịch vụ hoặc những người cần phải di chuyển qua khu vực có mật độ giao thông cao như Manhattan, có thể là những chuyến đi dài. Điều này có thể dẫn đến nhu cầu sử dụng dịch vụ đi lại như taxi hoặc ride-hailing tăng lên vào thời điểm này.
* Giá cước trung bình thấp nhất vào khoảng 8h sáng, thì đây là lúc mà tất cả người đi làm đều đang ở cơ quan làm việc, học sinh thì đi học ở trường, nên việc di chuyển bằng taxi có thể ít đi.
* Có thể thấy giá cước taxi trung bình theo giờ trong ngày có sự biến động liên tục.

3.4.3. Phân tích vị trí đón và trả khách  
 Trong phần này chúng ta sẽ tìm ra đâu là vị trí mà các chuyến taxi đón và trả khách nhiều nhất.



*Hình 20**. Code vẽ biểu đồ phân tán phân tích vị trí đón khách*

*Hình 21**. Biểu đồ phân tích vị trí taxi đón khách và giá cước*

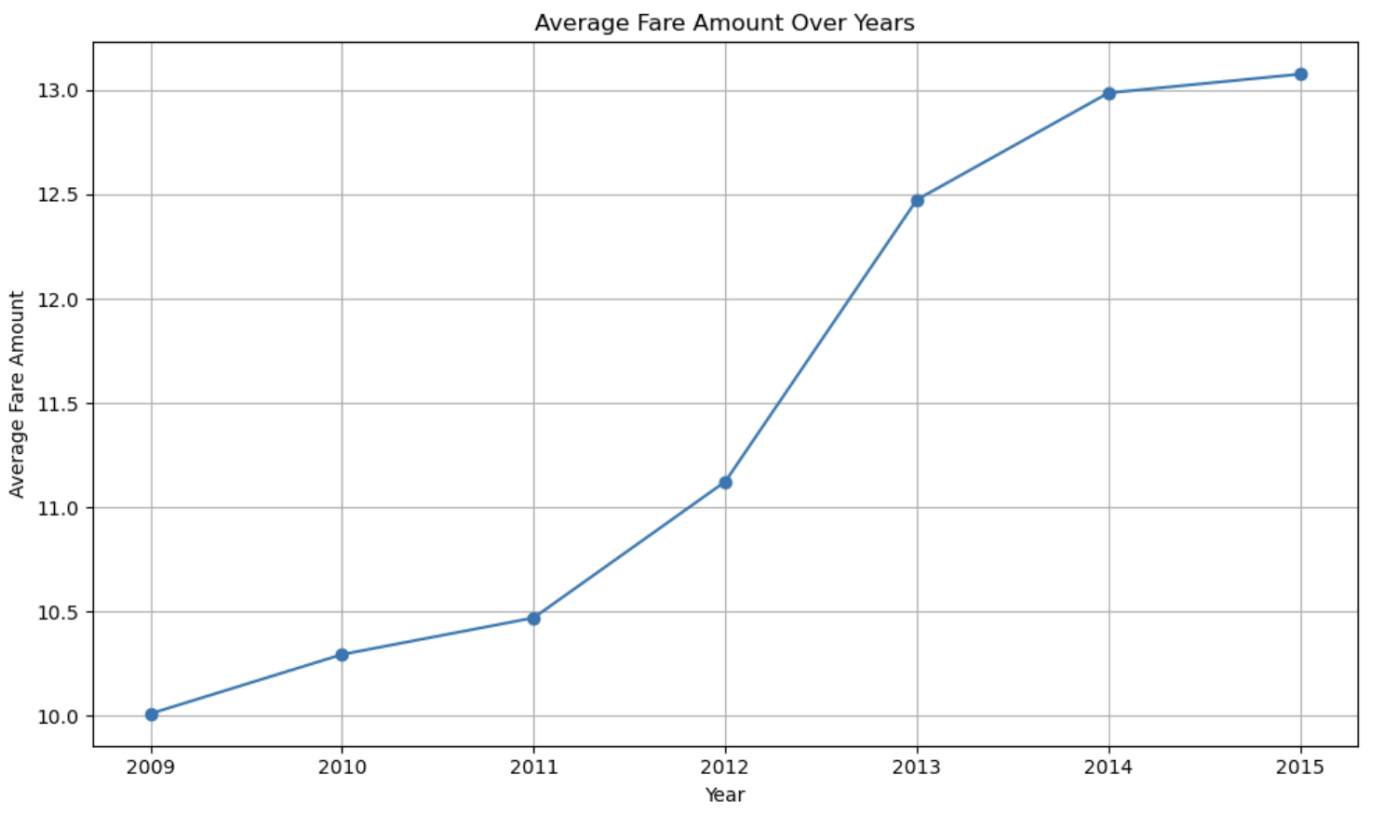
Dựa vào biểu đồ "Pickup Locations" ta có thể thấy:

* Khu vực đón khách chủ yếu ở New York City (NYC) tập trung chủ yếu ở vĩ độ khoảng 40.75 đến 41 và kinh độ từ -74.00 đến -73.75.
* Để biết được đó là những vị trí nào thì ta có thể tra trên Google Map.
* Trong đó có 1 nơi là tượng nữ thần tự do, có vị trí: 40.7187615,-74.1348721, và khi đối chiếu lên biểu đồ thì ta có thể thấy địa điểm này là nơi đón khách rất đông.
* Đa số các chuyến đi có giá cước dưới 50$.

#### 3.4.4. Phân tích xu hướng giá cước taxi qua từng năm



Hình 22. Code tạo biểu đồ đường phân tích xu hướng giá cước taxi qua từng năm



*Hình 23**. Biểu đồ trung bình giá cước taxi qua từng năm*

**Từ những nội dung phân tích bên trên, chúng ta rút ra được những thông tin sau:**

- Phần lớn các chuyến đi nằm trong mức giá cước dưới 50$.

- Giá cước cao nhất vào khoảng 5 giờ sáng, chủ yếu do các chuyến đi dài rời khỏi thành phố hoặc do người dân đi làm từ sớm.

- Trung bình giá cước taxi có mức cao nhất vào ngày chủ nhật.

- Sân bay và trung tâm thành phố, tượng nữ thần tự do là những điểm nóng mà taxi hay đến.

- Trung bình giá cước taxi có xu hướng tăng qua các năm từ 2019-2015 (tăng từ 10.0$ đến trên 13.0$).

**Kết luận từ việc phân tích tập dữ liệu:**

- Mức giá cước từ 50$ trở xuống là khá hợp lí đối với người dân ở New York City, bên cạnh đó điều này còn cho thấy rằng dịch vụ taxi là một lựa chọn phổ biến và hợp lý cho nhu cầu đi lại của người dân và du khách.

**Dự đoán tương lai:**

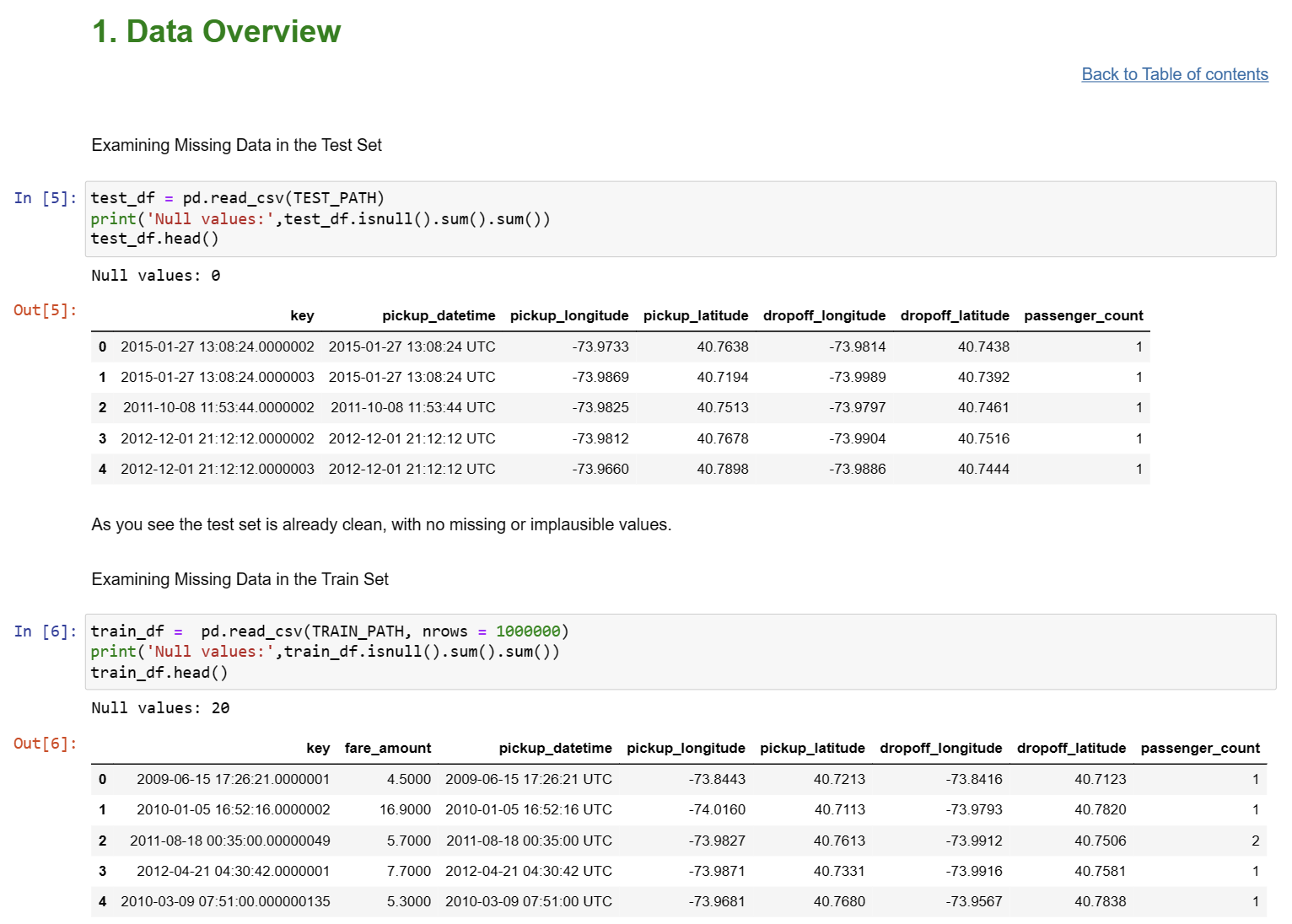
- Tăng giá cước vào khoảng 5 giờ sáng và vào cuối tuần: Với sự tăng trưởng kinh tế và tăng cường hoạt động của thành phố, dự đoán giá cước vào những thời điểm này sẽ tiếp tục tăng lên do nhu cầu đi lại tăng cao từ người dân đi làm từ sớm và từ các hoạt động giải trí, du lịch vào cuối tuần.

- Sân bay và trung tâm thành phố vẫn là điểm nóng của taxi: Với sự phát triển của ngành du lịch và tăng cường hoạt động kinh doanh, dự đoán rằng sân bay và trung tâm thành phố sẽ tiếp tục là những điểm tập trung lớn của người dân và du khách, là nơi mà taxi thường đến và đi.

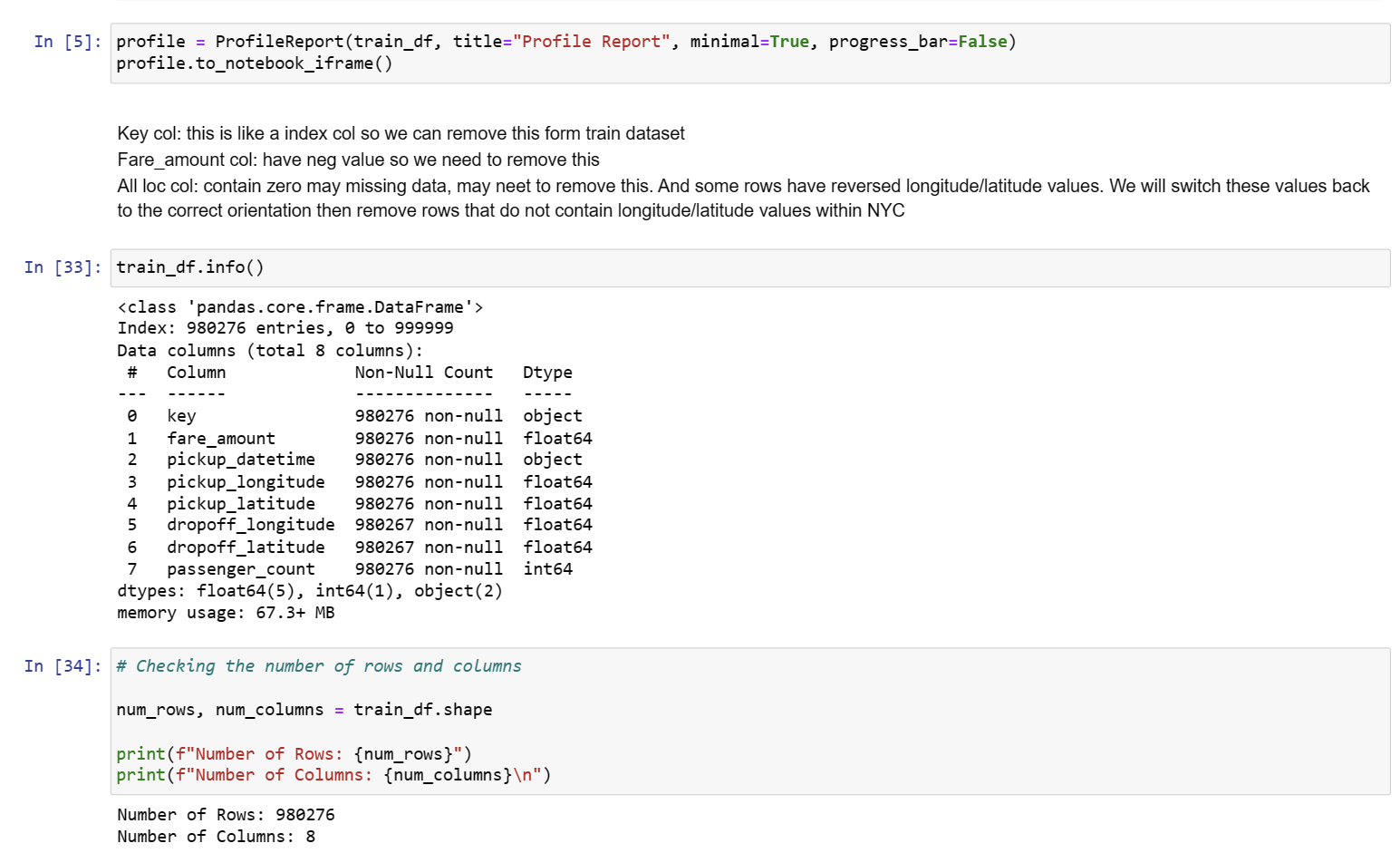
## CHƯƠNG 4: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN, KIỂM CHỨNG VÀ ĐÁNH GIÁ MÔ HÌNH CHO TẬP DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY

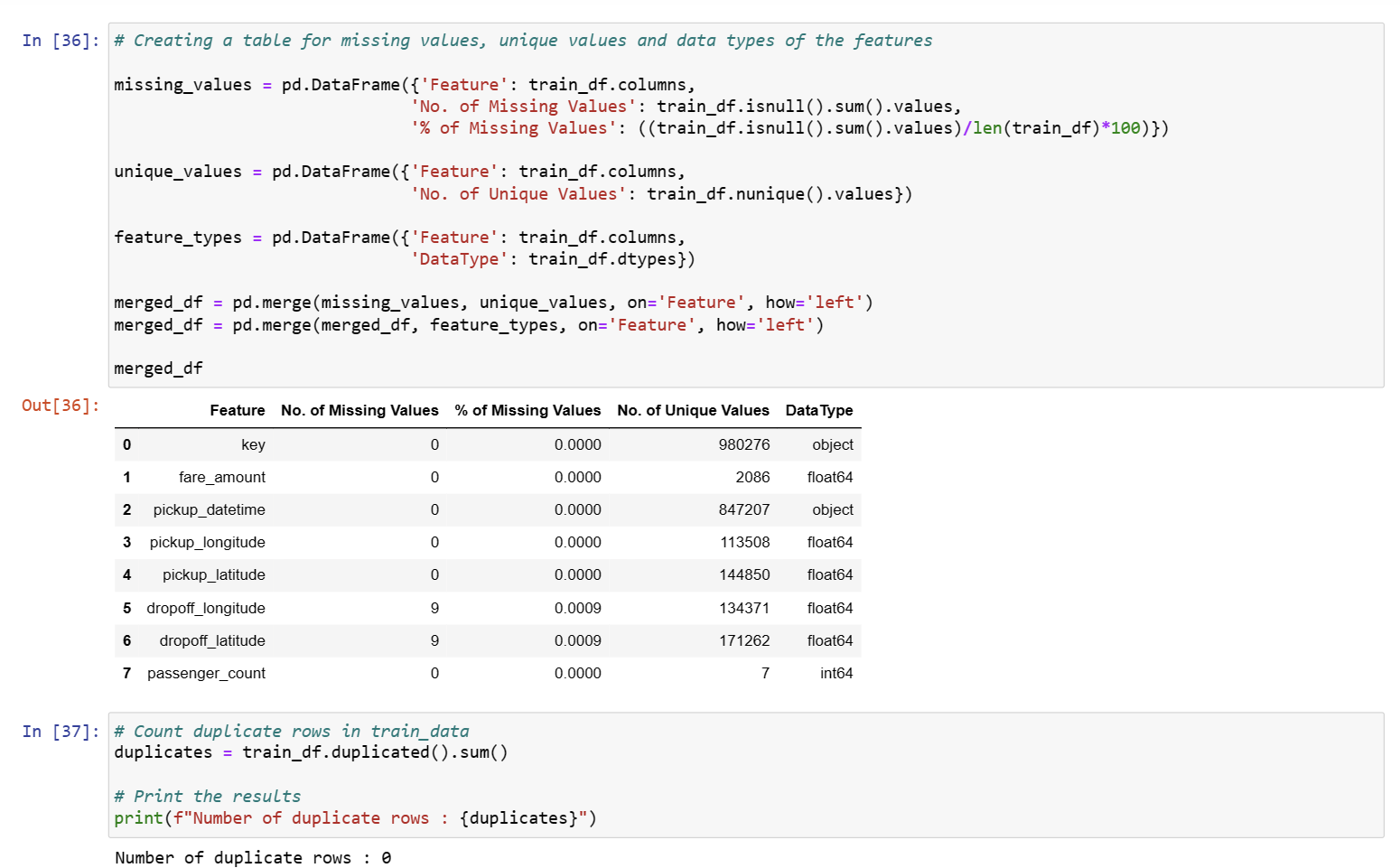
### 4.1. Tổng quan dữ liệu (Data Overview)

Kiểm nghiệm lại các tập dữ liệu train.csv, test.csv và clean data để thực hiện xây dựng mô hình dự đoán.

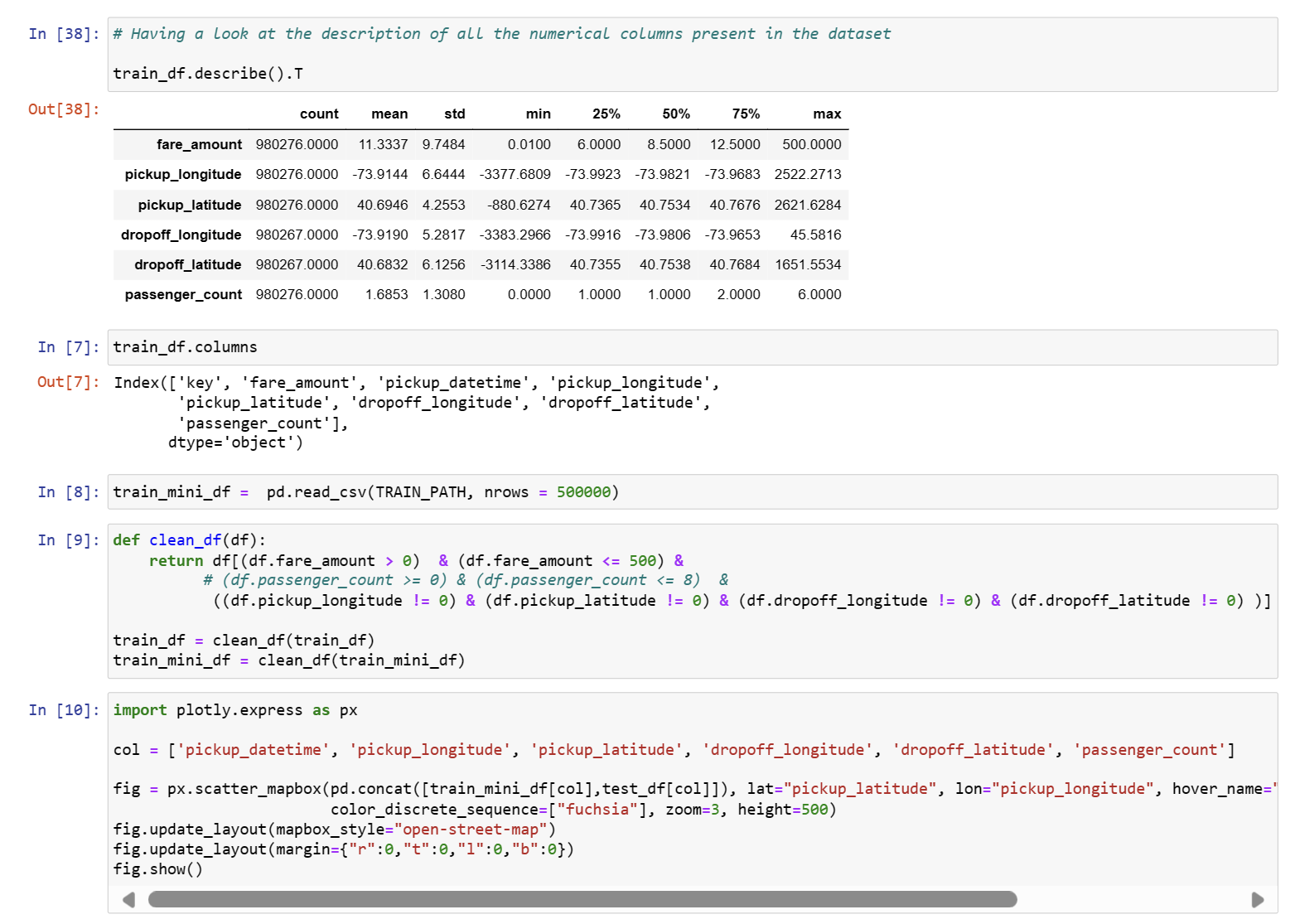


*Hình 24**. Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra và clean data*





*Hình 25**. Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra và clean data*



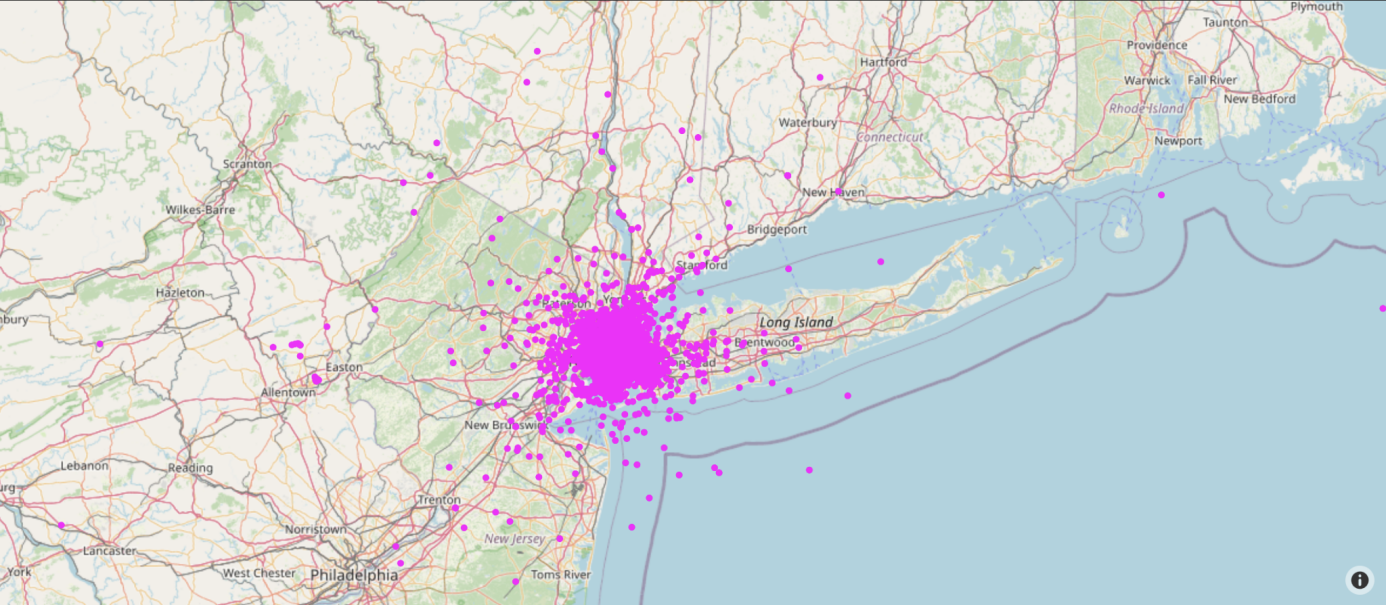
*Hình 26**. Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra và clean data*

### 4.2. Trực quan hóa vị trí taxi đón khách sử dụng plotly express

Trong phần này, chúng ta sẽ hiển thị các điểm đón khách của các chuyến xe taxi trên bản đồ sử dụng vĩ độ và kinh độ từ dữ liệu kết hợp của các tập dữ liệu huấn luyện và kiểm tra.



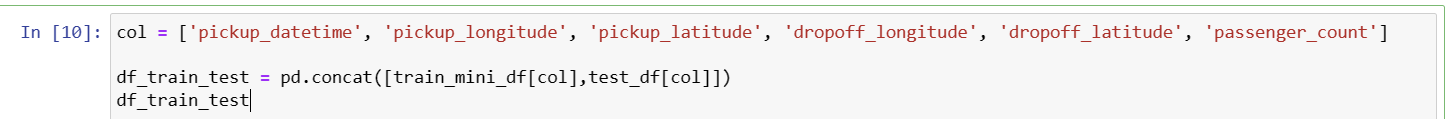
*Hình 27**. Code tạo bản đồ trực quan hóa các vị trí taxi đón khách ở New York City*



*Hình 28**. Bản đồ các vị trí mà taxi đón khách ở New York City*

### 4.3. Trực quan hóa các tuyến đường taxi sử dụng plotly graph objects

Trong phần này, chúng ta sẽ hiển thị các tuyến đường của các chuyến xe taxi trên bản đồ, sử dụng vĩ độ và kinh độ từ dữ liệu kết hợp của các tập dữ liệu huấn luyện và kiểm tra. Mỗi tuyến đường được biểu diễn bằng một đường thẳng nối giữa vị trí đón (pickup) và trả (dropoff) khách.



Hình 29. *Code kết hợp các cột liên quan từ hai DataFrame train và test*

*thành một DataFrame duy nhất*



*Hình 30**. Code tạo bản đồ trực quan hóa các tuyến đường taxi ở New York City*



*Hình 31**. Bản đồ các tuyến đường taxi ở New York City*

### 4.4. Tiền xử lý dữ liệu (Preprocessing Data)

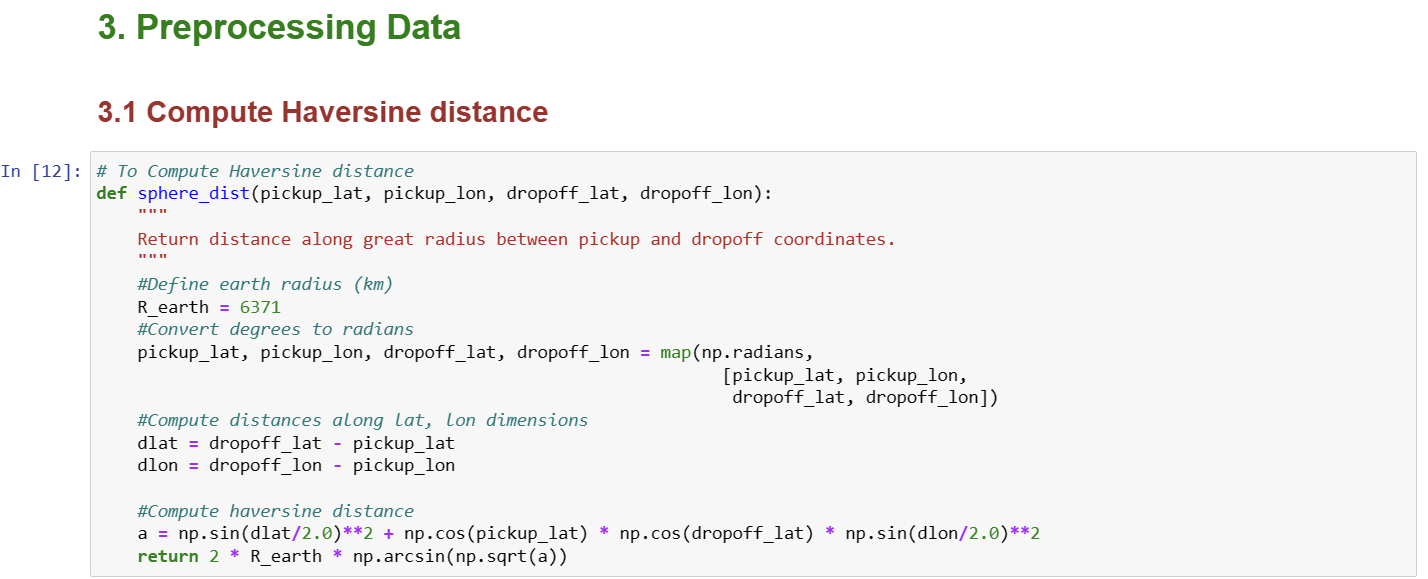
Chúng ta cần tạo ra các thông tin có khả năng và mối liên quan mật thiết nhất, có sự tác động cao đối với giá cước.

Trong trường hợp này khoản cách, vị trí phổ biến là các thông tin có mối tương quan cao và có thể đo dựa trên tập dữ liệu.

Cũng có nhũng thông tin khác có mối tương quan cao như quảng đường di chuyển, tốc độ, thời gian(cao điểm), các yếu tố kỹ thuật, điều kiện môi trường,…

#### 4.4.1. Tính khoảng cách Haversine (Compute Haversine distance)

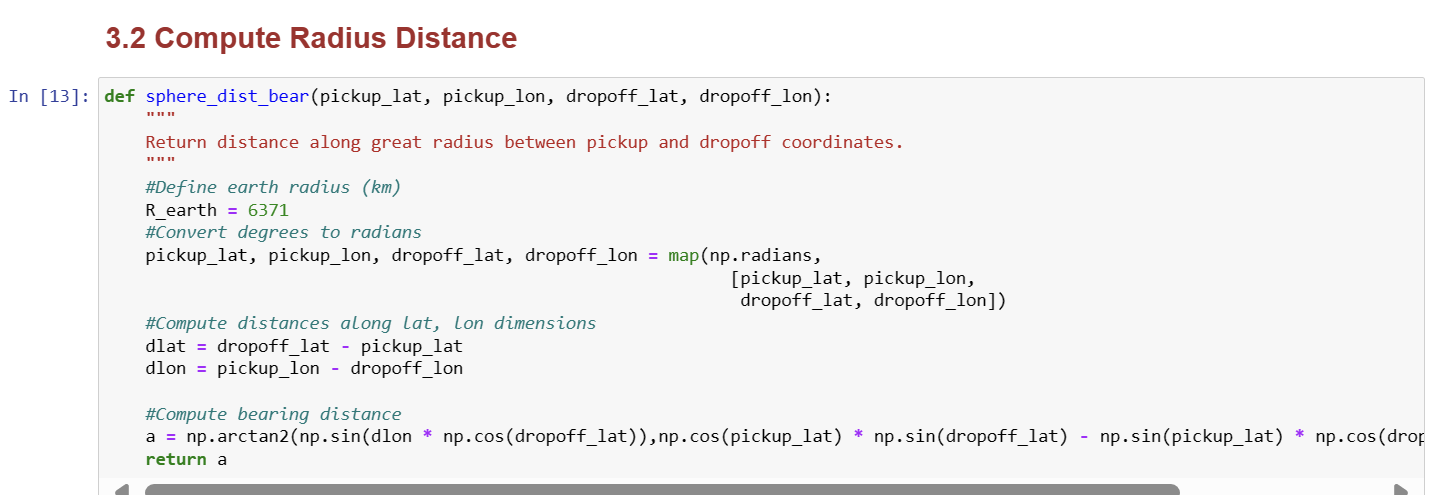
Tạo hàm để tính khoảng cách Haversine giữa hai điểm trên bề mặt Trái Đất, sử dụng tọa độ vĩ độ và kinh độ của điểm đón và điểm trả khách.



*Hình 32**. Code tạo hàm tính khoảng cách Haversine*

#### 4.4.2. Tính khoảng cách radius (Compute Radius Distance)

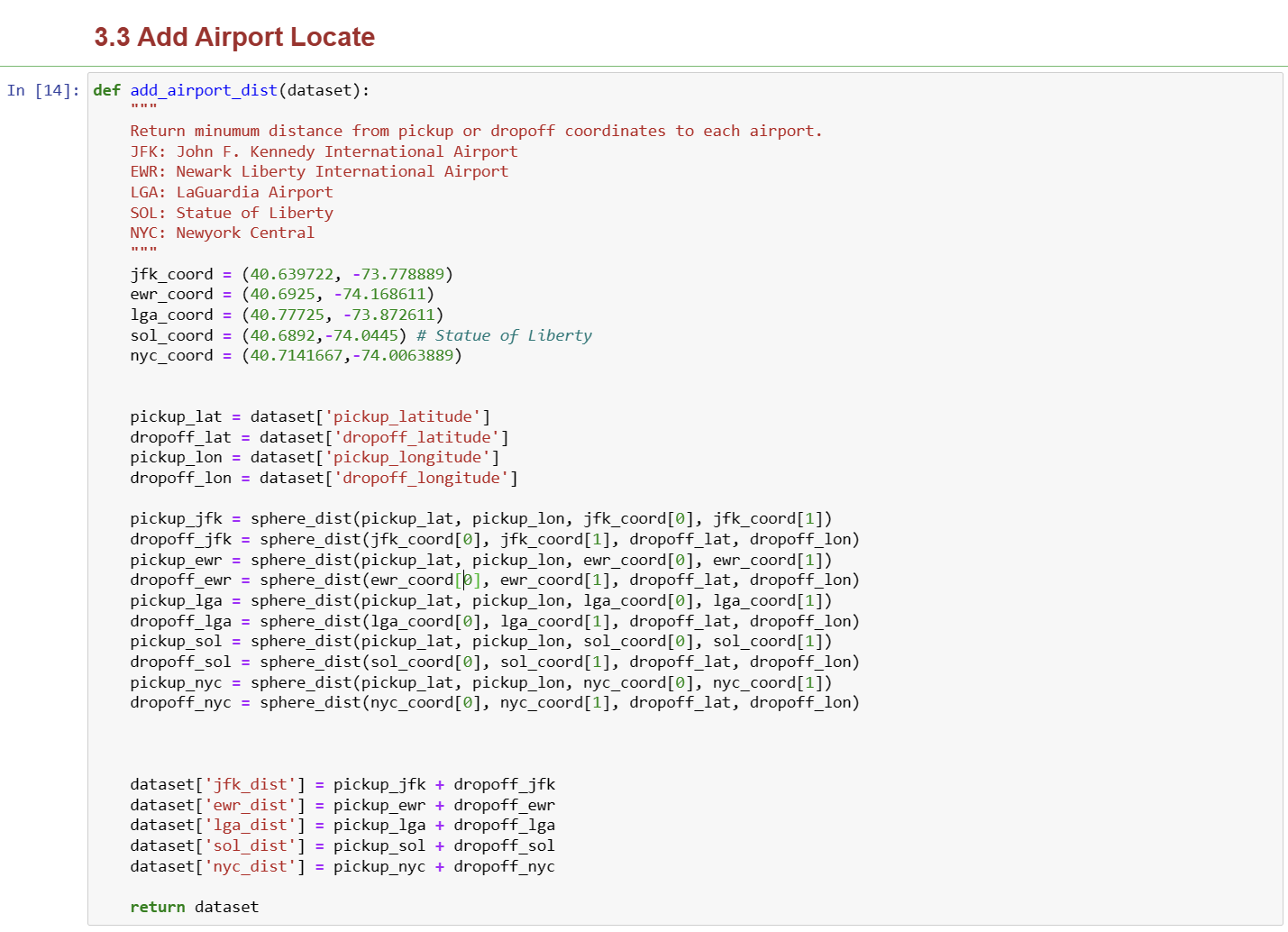
Tạo hàm để tính toán khoảng cách theo hướng (bearing) giữa hai điểm, sử dụng tọa độ vĩ độ và kinh độ của điểm đón và điểm trả khách.



*Hình 33**. Code tạo hàm để tính toán hướng (bearing)*

#### 4.4.3. Tính khoảng cách đến các địa điểm nổi bật ở New York (Add Airport Locate)

Tạo hàm để tính toán khoảng cách từ điểm đón và trả khách của mỗi chuyến taxi đến năm địa điểm quan trọng (JFK, EWR, LGA *(mã của ba sân bay quốc tế chính ở New York)*, Tượng Nữ thần Tự do, và trung tâm New York) và thêm các khoảng cách này vào DataFrame dưới dạng các cột mới. Mỗi cột mới đại diện cho tổng khoảng cách từ điểm đón và trả khách đến một địa điểm cụ thể.

**

*Hình 34**. Code tạo hàm tính khoảng cách đến các địa điểm nổi bật ở New York City*

4.4.4. Tiền xử lý (Preprocess)

Tạo các hàm tiền xử lý dữ liệu:

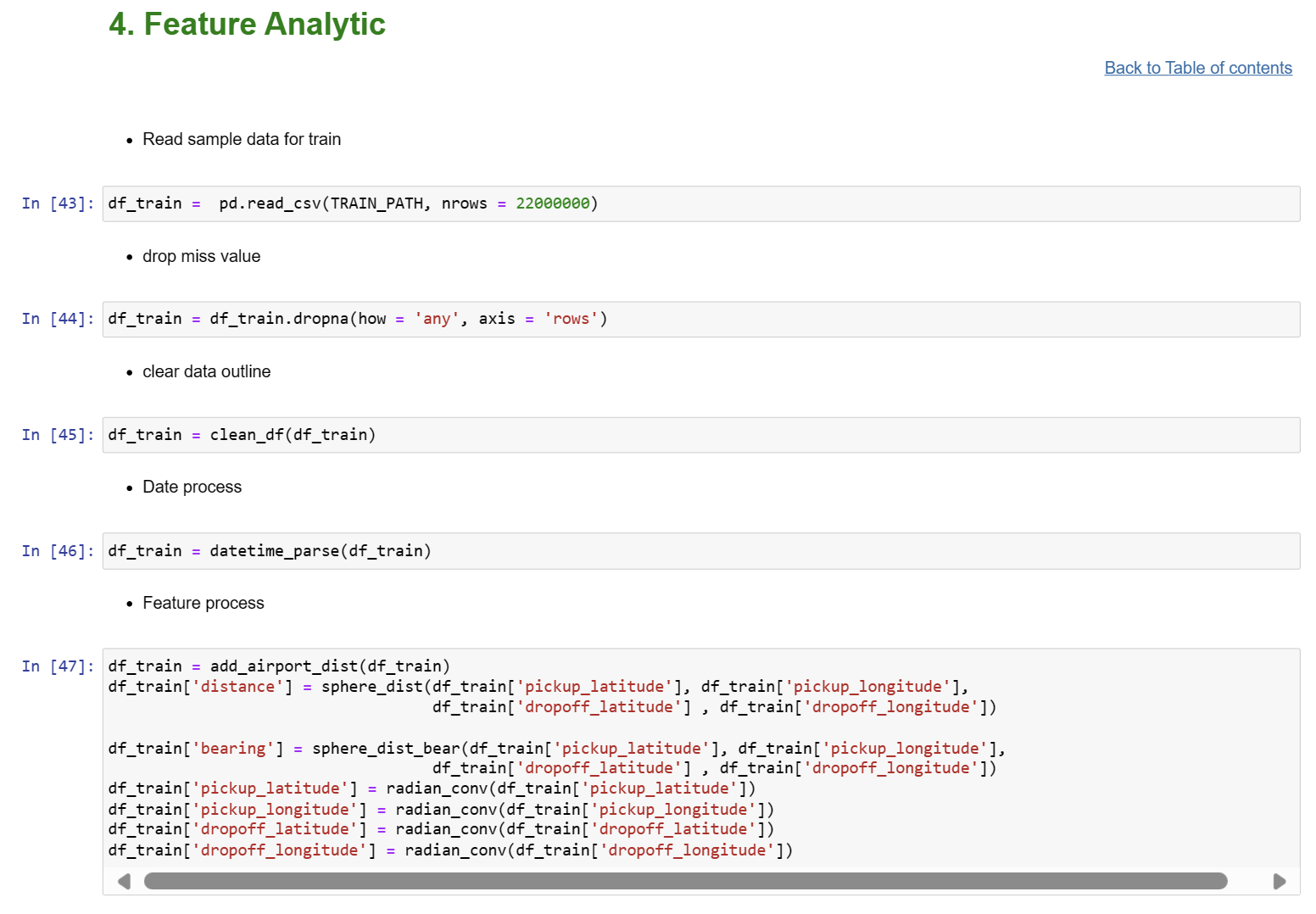
* Hàm radian\_conv: Chuyển đổi độ sang radian.
* Hàm datetime\_parse: Chuyển đổi cột thời gian sang định dạng datetime và trích xuất thông tin thời gian.
* Hàm clean\_df: Làm sạch dữ liệu bằng cách loại bỏ các bản ghi không hợp lệ hoặc nằm ngoài phạm vi địa lý được xác định.



*Hình 35**. Code tạo các hàm tiền xử lý dữ liệu*

### 4.4. Phân tích đặc trưng (Feature Analytic)

Phân tích và tiền xử lý dữ liệu để chuẩn bị cho việc tạo ra các đặc trưng (features) để sử dụng trong mô hình dự đoán giá cước taxi.



*Hình 36**. Code thực hiện phân tích và tiền xử lý dữ liệu*

### 4.5. Xây dựng Model (Build Model)

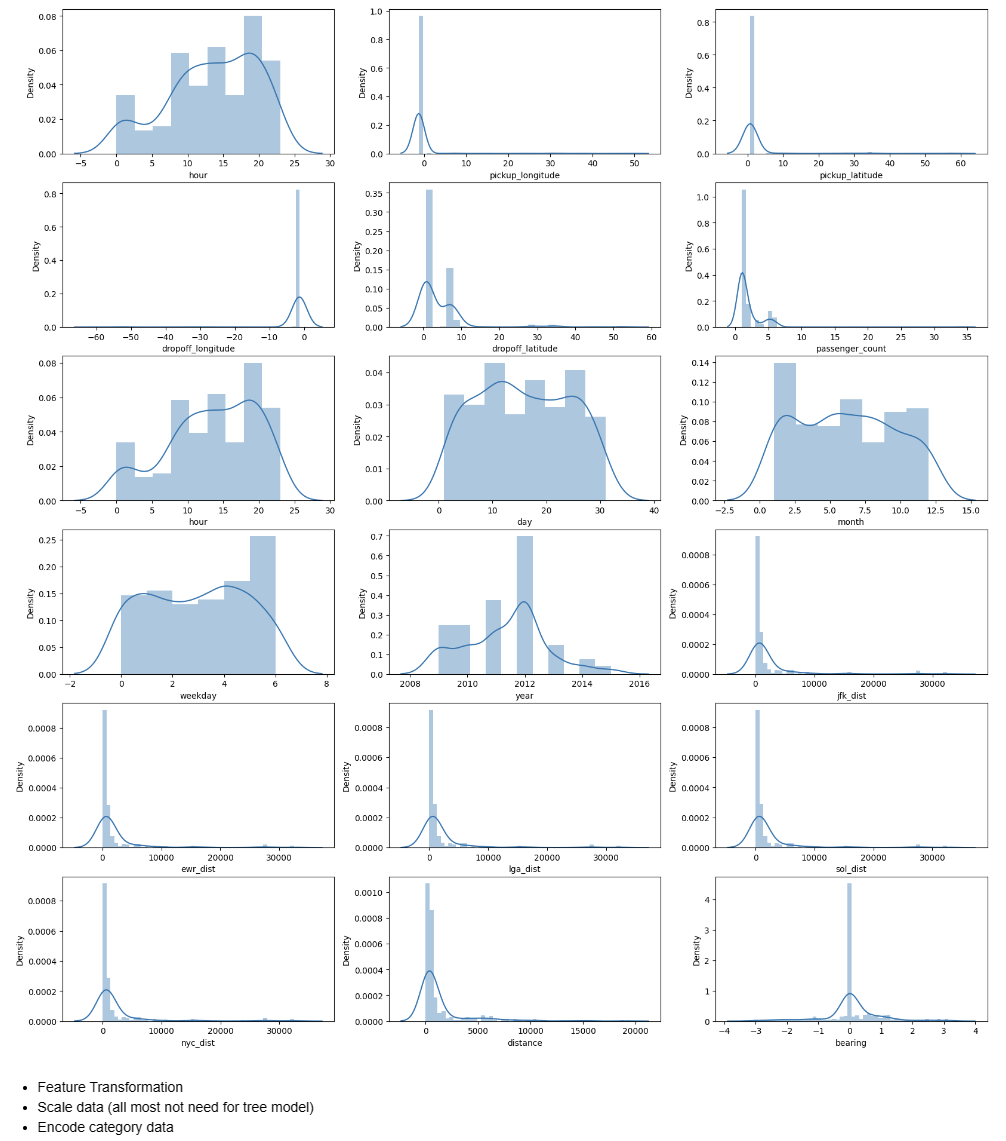
#### 4.5.1. Train-Val split

Trước khi xây dựng mô hình dự đoán giá cước taxi, chúng ta cần chia dữ liệu thành tập huấn luyện và tập validation để đánh giá hiệu suất của mô hình. Tiếp theo, chúng ta sẽ vẽ biểu đồ phân phối của các đặc trưng trong tập huấn luyện để hiểu rõ hơn về phân bố của chúng. Điều này sẽ giúp chúng ta đưa ra quyết định thông minh hơn khi xây dựng và điều chỉnh mô hình dự đoán giá cước taxi.

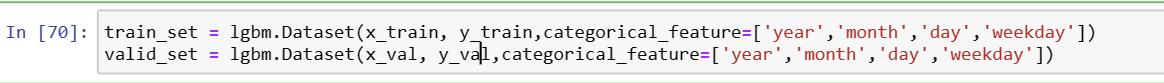




*Hình 37**. Code trực quan hóa phân phối đặc trưng trong dữ liệu huấn luyện*



*Hình 38**. Kết quả trực quan hóa phân phối đặc trưng trong dữ liệu huấn luyện*

**

*Hình 39**. Code chuẩn bị dữ liệu huấn luyện và validation cho mô hình LightGBM*

#### 4.5.2. Train Model

Trong phần này chúng ta sẽ thực hiện việc huấn luyện mô hình sử dụng thuật toán Gradient Boosting Decision Tree (GBDT) thông qua thư viện LightGBM.



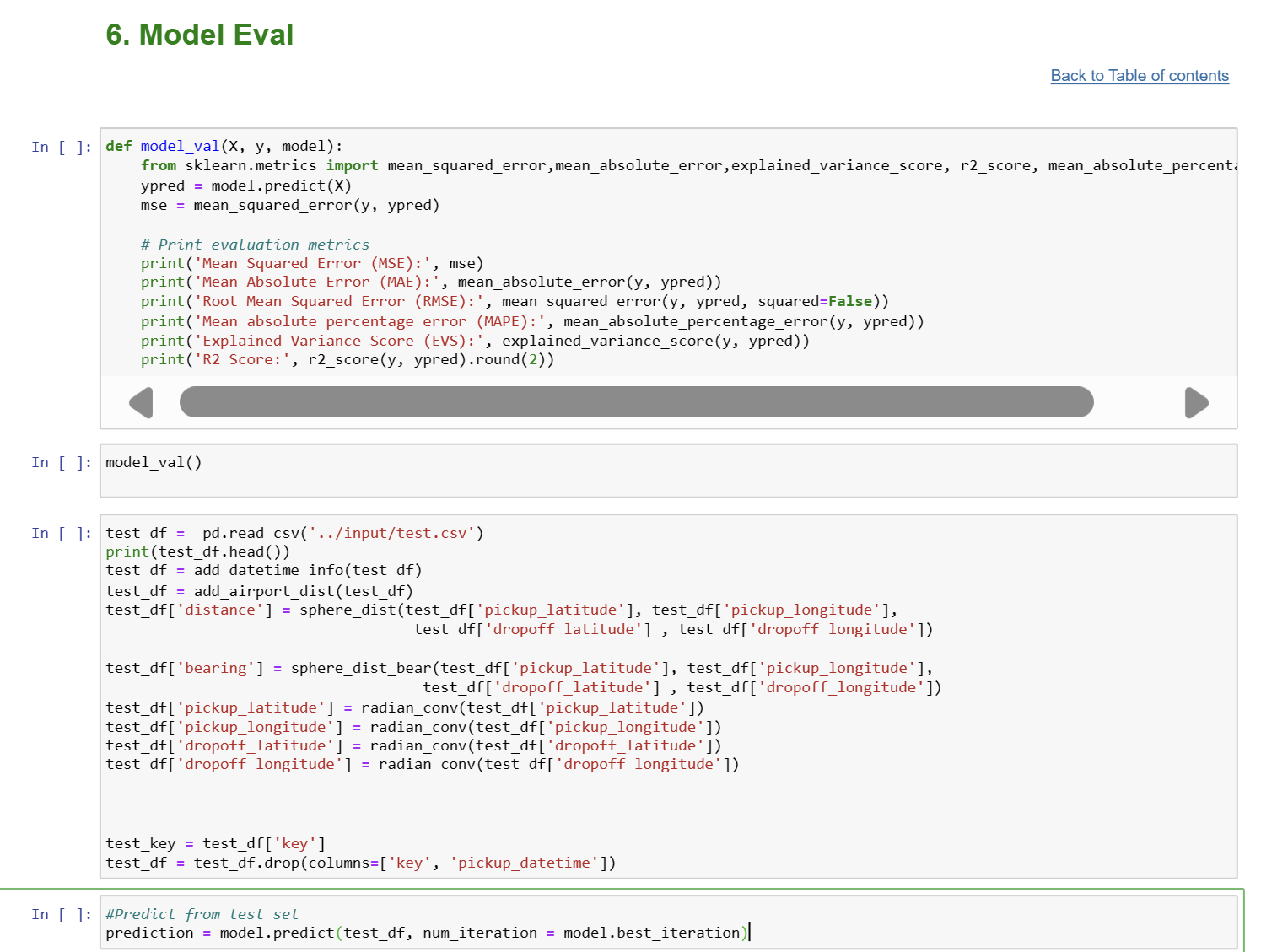
A screenshot of a computer program

Description automatically generated

*Hình 40**. Code huấn luyện mô hình*

### 4.6. Mô hình Eval (Model Eval)

Ở phần này chúng ta sẽ tiến hành đánh giá hiệu suất của mô hình dự đoán giá cước taxi. Bằng cách sử dụng các độ đo như Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Explained Variance Score (EVS), và R2 Score. Sau đó chúng ta sẽ dự đoán giá cước taxi cho tập dữ liệu kiểm tra.



*Kêt quả đánh giá:*

*A screenshot of a computer

Description automatically generated*

*A screenshot of a computer code

Description automatically generated*

*Hình 41**. đánh giá hiệu suất mô hình*

# PHẦN KẾT LUẬN

* + - 1. **Kết quả đạt được**
* Về kiến thức

+ Đựa trên tập dữ liệu:

Kiến thức về phần tích dữ liệu lớn, đến rất lớn. Tập trung hiểu được cốt lõi các giá trị thành phần chính quan trọng của dữ liệu, cách chúng tác động tới nhau.

Bằng các bước tìm hiểu trên tập dữ liệu lớn giúp chúng em: Bước đầu nắm bắt được những lỗi trong dữ liệu và trong quá trình thu thập. Kế đến các bước xử lí bằng sample, các tập con nhỏ, có tính khái quát cao, và các tập lớn hơn một cách hơp lý.

Tốc độ tính toán trên tập dữ liệu lớn khá chậm, nên cần có các phương pháp xử lý giúp cải thiện tốc độ. Mang lại hiệu qua bao phủ cao.

+ Trên quá trình train model:

Hiểu dữ liệu là một bước vô cùng quan trọn, từ đó chúng em có khả năng đưa ra được các phân, lực chọn các đặt trương hợp lý, có mới tương quan cao.

Hiểu được việc xử lý dữ liệu ảnh hưởng đến kết quả model như thế nào.

Biết được các tối ưu các siêu tham số để tạo được tập model tốt nhất.

* Về sản phẩm

+ Sản phẩm có khả năng dự đoán tốt giá cả hợp lý, độ chính xác để đề xuất mức giá dự đoán trước cao.

+ Tính ứng dụng model cho các mô hình báo giá taxi tốt, có thể làm mức giá quy chuẩn cho các doanh nghiêp taxi

* + - 1. **Ưu và nhược điểm**
* Ưu điểm: Mô hình dự đoán nhanh, mạnh, tính chính xác cao có khả năng ứng dụng triển khai trên các sản phẩm di động.
* Nhược điểm: Độ bao phủ thấp (tập trung ở new york), chưa có sự linh hoạt do thiếu các đặt trưng về môi trường, chi phí, điều kiện vận hành.
  + - 1. **Hướng phát triển**
* Fine turning trên nhiều tập dataset của nhiều khu vực hơn.
* Thêm các yếu tố môi trường cho sản phẩm.

# PHẦN TÀI LIỆU THAM KHẢO

[1]. Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow,* Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2019

[2]. Kaggle (2018), *New York City Taxi Fare Prediction*, truy cập ngày 10/5/2024, đường dẫn: <https://www.kaggle.com/competitions/new-york-city-taxi-fare-prediction/data>